

Docket No.: 32178-201736  
(PATENT)

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

In re Patent Application of:  
Okumura et al.

Application No.: Not Yet Assigned

Confirmation No.: Not Yet Assigned

Filed: Concurrently Herewith

Art Unit: N/A

For: INFORMATION EXTRACTING APPARATUS

Examiner: Not Yet Assigned

**CLAIM FOR PRIORITY AND SUBMISSION OF DOCUMENT**

MS Patent Application  
Commissioner for Patents  
P.O. Box 1450  
Alexandria, VA 22313-1450

Dear Sir:


Applicant hereby claims priority under 35 U.S.C. 119 based on the following prior foreign application filed in the following foreign country on the date indicated:

<u>Country</u>	<u>Application No.</u>	<u>Date</u>
Japan	2003-098165	April 1, 2003

In support of this claim, a certified copy of the said original foreign application is filed herewith.

Dated: March 25, 2004

Respectfully submitted,

By   
Michael A. Sartori, Ph.D.  
Registration No.: 41,289  
VENABLE LLP  
P.O. Box 34385  
Washington, DC 20043-9998  
(202) 344-4000  
(202) 344-8300 (Fax)  
Attorney/Agent For Applicant

日本国特許庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 2003年 4月 1日  
Date of Application:

出願番号 特願2003-098165  
Application Number:  
[ST. 10/C]: [JP 2003-098165]

出願人 沖電気工業株式会社  
Applicant(s):

2003年12月15日

特許庁長官  
Commissioner,  
Japan Patent Office

今井 康



出証番号 出証特2003-3104014

【書類名】 特許願

【整理番号】 SA003803

【あて先】 特許庁長官殿

【国際特許分類】 G06F 17/30

【発明者】

    【住所又は居所】 東京都港区虎ノ門 1 丁目 7 番 1 2 号 沖電気工業株式会  
社内

    【氏名】 奥村 晃弘

【発明者】

    【住所又は居所】 東京都港区虎ノ門 1 丁目 7 番 1 2 号 沖電気工業株式会  
社内

    【氏名】 大沼 宏行

【発明者】

    【住所又は居所】 東京都港区虎ノ門 1 丁目 7 番 1 2 号 沖電気工業株式会  
社内

    【氏名】 濱口 佳孝

【特許出願人】

    【識別番号】 000000295

    【氏名又は名称】 沖電気工業株式会社

【代理人】

    【識別番号】 100082050

    【弁理士】

    【氏名又は名称】 佐藤 幸男

【手数料の表示】

    【予納台帳番号】 058104

    【納付金額】 21,000円

【提出物件の目録】

    【物件名】 明細書 1

    【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9100477

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 情報抽出装置

【特許請求の範囲】

【請求項 1】 リンク情報で相互に関連付けられたハイパーテキスト構造を持つ文書群から指定した情報を抽出する情報抽出装置であって、

前記情報を抽出する起点となる文書のアドレスを指定する起点アドレス指定部と、

前記起点アドレス指定部で指定された対象文書から前記情報を抽出すると共に、前記対象文書から当該情報を抽出できなかった場合は、前記文書のアドレスに基づいて前記対象文書の関連文書から当該情報を抽出する抽出部とを備えたことを特徴とする情報抽出装置。

【請求項 2】 請求項 1 に記載の情報抽出装置において、

抽出する情報のカテゴリを指定するカテゴリ指定部と、

起点アドレス指定部で指定された対象文書から前記カテゴリに該当する情報を抽出すると共に、前記対象文書から当該カテゴリに該当する情報を抽出できなかった場合は、前記文書のアドレスに基づいて前記対象文書の関連文書から当該情報を抽出する抽出部とを備えたことを特徴とする情報抽出装置。

【請求項 3】 請求項 2 に記載の情報抽出装置において、

抽出する情報のカテゴリを階層構造で表現したカテゴリ階層規定部と、

起点アドレス指定部で指定された対象文書からカテゴリに該当する情報を抽出した結果、前記階層構造のうち下位階層の抽出結果のみで上位階層の抽出結果が欠落している場合は、対象文書の関連文書から前記下位階層の抽出結果よりも上位階層の文字列を抽出する抽出部と、

前記下位階層の抽出結果と前記上位階層の抽出結果を合成した文字列を抽出結果として出力する処理部とを備えたことを特徴とする情報抽出装置。

【請求項 4】 請求項 3 に記載の情報抽出装置において、

起点アドレス指定部で指定された対象文書からカテゴリに該当する情報を抽出した結果、前記階層構造のうち下位階層の抽出結果と上位階層の抽出結果の複数の文字列に分かれた場合は、これら複数の文字列を、下位階層の抽出結果と上位

階層の抽出結果として出力する抽出部を備えたことを特徴とする情報抽出装置。

【請求項 5】 リンク情報で相互に関連付けられたハイパーテキスト構造を持つ文書群から指定した情報を抽出する情報抽出装置であって、

前記文書群から対象となる情報を抽出すると共に、前記文書群に対して文書の追加または更新が発生した場合は、その都度これを反映させた抽出処理を行い、前記対象となる情報とその文書アドレスとを含む抽出結果を出力する抽出部と、

前記抽出部からの抽出結果を抽出結果情報として記憶する抽出結果記憶部と、

前記指定した情報を抽出する起点となる文書のアドレスを指定する起点アドレス指定部と、

前記抽出結果記憶部の抽出結果情報を参照して、前記起点アドレス指定部で指定された文書アドレスの文書および関連文書から情報抽出を行う探索部とを備えたことを特徴とする情報抽出装置。

【請求項 6】 請求項 5 に記載の情報抽出装置において、

抽出を行う情報のカテゴリを指定するカテゴリ指定部と、

前記カテゴリ指定部で指定されたカテゴリに属する情報抽出を行う探索部とを備えたことを特徴とする情報抽出装置。

【請求項 7】 請求項 6 に記載の情報抽出装置において、

抽出する情報のカテゴリを階層構造で表現したカテゴリ階層規定部と、

起点アドレス指定部で指定された対象文書からカテゴリに該当する情報を抽出した結果、前記階層構造のうち下位階層の抽出結果のみで上位階層の抽出結果が欠落している場合は、対象文書の関連文書から前記下位階層の抽出結果よりも上位階層の文字列を抽出し、前記下位階層の抽出結果と前記上位階層の抽出結果を合成した文字列を抽出結果として出力する探索部とを備えたことを特徴とする情報抽出装置。

【請求項 8】 請求項 1 ～ 7 のいずれかに記載の情報抽出装置において、

関連文書は、対象文書のリンク先文書、リンク元文書、上位文書のうち、少なくともいずれか一つを含むことを特徴とする情報抽出装置。

【請求項 9】 請求項 8 に記載の情報抽出装置において、

上位文書は、対象文書の一つ上のディレクトリに存在する特定の名称の文書、

または、一つ上のディレクトリに存在するリンク元文書のうち、少なくともいずれかの文書であることを特徴とする情報抽出装置。

【請求項 10】 請求項 1～4 のいずれかに記載の情報抽出装置において、  
最大リンク深度を指定する最大リンク深度指定部と、  
対象文書から情報抽出できなかった場合は、その文書の関連文書から情報抽出を行う処理を、前記指定された最大リンク深度の範囲内で再帰的に行う抽出部とを備えたことを特徴とする情報抽出装置。

【請求項 11】 請求項 5～7 のいずれかに記載の情報抽出装置において、  
最大リンク深度を指定する最大リンク深度指定部と、  
対象文書から情報抽出できなかった場合は、その文書の関連文書から情報抽出を行う処理を、前記指定された最大リンク深度の範囲内で再帰的に行う探索部とを備えたことを特徴とする情報抽出装置。

【請求項 12】 請求項 10 に記載の情報抽出装置において、  
リンク深度の値が小さい文書から順に情報抽出処理を行う抽出部を備えたことを特徴とする情報抽出装置。

【請求項 13】 請求項 11 に記載の情報抽出装置において、  
リンク深度の値が小さい文書から順に情報抽出処理を行う探索部を備えたことを特徴とする情報抽出装置。

【請求項 14】 請求項 1～4、10、12 のいずれかに記載の情報抽出装置において、

関連文書の文書アドレスに基づいて内部リンクと外部リンクとを判別し、外部リンクの文書は情報抽出の対象から除外する抽出部を備えたことを特徴とする情報抽出装置。

【請求項 15】 請求項 5～7、11、13 のいずれかに記載の情報抽出装置において、

関連文書の文書アドレスに基づいて内部リンクと外部リンクとを判別し、外部リンクの文書は情報抽出の対象から除外する探索部を備えたことを特徴とする情報抽出装置。

【請求項 16】 請求項 3 または 4 に記載の情報抽出装置において、

階層構造に基づいて、上位階層の抽出結果から下位階層の抽出結果の順番に複数の文字列を結合することにより処理結果の文字列を作成する処理部を備えたことを特徴とする情報抽出装置。

【請求項 17】 請求項 7 に記載の情報抽出装置において、

階層構造に基づいて、上位階層の抽出結果から下位階層の抽出結果の順番に複数の文字列を結合することにより処理結果の文字列を作成する探索部を備えたことを特徴とする情報抽出装置。

【請求項 18】 請求項 3、4、16 に記載の情報抽出装置において、

階層構造で表現された複数の文字列を合成する場合の所定の合成ルールを有し、当該合成ルールに従って処理結果の文字列を作成する処理部を備えたことを特徴とする情報抽出装置。

【請求項 19】 請求項 7 または 17 に記載の情報抽出装置において、

階層構造で表現された複数の文字列を合成する場合の所定の合成ルールを有し、当該合成ルールに従って処理結果の文字列を作成する探索部を備えたことを特徴とする情報抽出装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、自然言語処理システムに関し、特に、特定の情報を抽出する情報抽出装置に関する。

【0002】

【従来の技術】

従来、特定の情報を抽出する情報抽出を用いた質問応答システムがあった（例えば、特許文献 1 参照）。このような質問応答システムとは、文書集合と質問文が与えられると、その質問文に対する回答を出力するシステムである。このシステムでは、入力された質問文から検索語集合と質問種別を判定し、その検索語集合および質問種別に従って、与えられた文書集合から関連文書集合を検索し、その関連文書集合の各文書から回答を抽出して出力する。この検索した文書集合から回答を抽出する部分に情報抽出が用いられている。



## 【0003】

## 【特許文献1】

特開 2002-132811 号公報

## 【0004】

## 【発明が解決しようとする課題】

上記従来の質問応答システムにおける情報抽出では、システムに与える文書集合がハイパーテキスト形式で記述された文書である場合については特に示されていない。しかしながら、ハイパーテキスト形式で記述された文書は、本来一つの文書にする筈のものを、読み易さを向上させるために複数に分割し、それらを互いにリンクさせている場合がある。このような場合、検索した文書からのみ情報を抽出するだけでは不十分であり、検索した文書のリンク先の文書からも抽出する必要があった。

## 【0005】

特に、近年はインターネットの発達もあって、ハイパーテキスト形式で記述された文書が非常に増えてきている。このため、これらの文書を的確に処理できないことは質問応答システムだけでなく、情報抽出を用いる種々のシステムにとっても大きな問題となっていた。

## 【0006】

## 【課題を解決するための手段】

本発明は、前述の課題を解決するため次の構成を採用する。

## 〈構成1〉

リンク情報で相互に関連付けられたハイパーテキスト構造を持つ文書群から指定した情報を抽出する情報抽出装置であって、情報を抽出する起点となる文書のアドレスを指定する起点アドレス指定部と、起点アドレス指定部で指定された対象文書から情報を抽出すると共に、対象文書から情報を抽出できなかった場合は、文書のアドレスに基づいて対象文書の関連文書から情報を抽出する抽出部とを備えたことを特徴とする情報抽出装置。

## 【0007】

## 〈構成2〉

構成 1 に記載の情報抽出装置において、抽出する情報のカテゴリを指定するカテゴリ指定部と、起点アドレス指定部で指定された対象文書からカテゴリに該当する情報を抽出すると共に、対象文書からカテゴリに該当する情報を抽出できなかった場合は、文書のアドレスに基づいて対象文書の関連文書から情報を抽出する抽出部とを備えたことを特徴とする情報抽出装置。

#### 【0008】

##### 〈構成 3〉

構成 2 に記載の情報抽出装置において、抽出する情報のカテゴリを階層構造で表現したカテゴリ階層規定部と、起点アドレス指定部で指定された対象文書からカテゴリに該当する情報を抽出した結果、階層構造のうち下位階層の抽出結果のみで上位階層の抽出結果が欠落している場合は、対象文書の関連文書から下位階層の抽出結果よりも上位階層の文字列を抽出する抽出部と、下位階層の抽出結果と上位階層の抽出結果を合成した文字列を抽出結果として出力する処理部とを備えたことを特徴とする情報抽出装置。

#### 【0009】

##### 〈構成 4〉

構成 3 に記載の情報抽出装置において、起点アドレス指定部で指定された対象文書からカテゴリに該当する情報を抽出した結果、階層構造のうち下位階層の抽出結果と上位階層の抽出結果の複数の文字列に分かれた場合は、これら複数の文字列を、下位階層の抽出結果と上位階層の抽出結果として出力する抽出部を備えたことを特徴とする情報抽出装置。

#### 【0010】

##### 〈構成 5〉

リンク情報で相互に関連付けられたハイパーテキスト構造を持つ文書群から指定した情報を抽出する情報抽出装置であって、文書群から対象となる情報を抽出すると共に、文書群に対して文書の追加または更新が発生した場合は、その都度これを反映させた抽出処理を行い、対象となる情報とその文書アドレスとを含む抽出結果を出力する抽出部と、抽出部からの抽出結果を抽出結果情報として記憶する抽出結果記憶部と、指定した情報を抽出する起点となる文書のアドレスを指

定する起点アドレス指定部と、抽出結果記憶部の抽出結果情報を参照して、起点アドレス指定部で指定された文書アドレスの文書および関連文書から情報抽出を行う探索部とを備えたことを特徴とする情報抽出装置。

#### 【0011】

##### 〈構成6〉

構成5に記載の情報抽出装置において、抽出を行う情報のカテゴリを指定するカテゴリ指定部と、カテゴリ指定部で指定されたカテゴリに属する情報抽出を行う探索部とを備えたことを特徴とする情報抽出装置。

#### 【0012】

##### 〈構成7〉

構成6に記載の情報抽出装置において、抽出する情報のカテゴリを階層構造で表現したカテゴリ階層規定部と、起点アドレス指定部で指定された対象文書からカテゴリに該当する情報を抽出した結果、階層構造のうち下位階層の抽出結果のみで上位階層の抽出結果が欠落している場合は、対象文書の関連文書から下位階層の抽出結果よりも上位階層の文字列を抽出し、下位階層の抽出結果と上位階層の抽出結果を合成した文字列を抽出結果として出力する探索部とを備えたことを特徴とする情報抽出装置。

#### 【0013】

##### 〈構成8〉

構成1～7のいずれかに記載の情報抽出装置において、関連文書は、対象文書のリンク先文書、リンク元文書、上位文書のうち、少なくともいずれか一つを含むことを特徴とする情報抽出装置。

#### 【0014】

##### 〈構成9〉

構成8に記載の情報抽出装置において、上位文書は、対象文書の一つ上のディレクトリに存在する特定の名称の文書、または、一つ上のディレクトリに存在するリンク元文書のうち、少なくともいずれかの文書であることを特徴とする情報抽出装置。

#### 【0015】

〈構成 10〉

構成 1～4 のいずれかに記載の情報抽出装置において、最大リンク深度を指定する最大リンク深度指定部と、対象文書から情報抽出できなかった場合は、その文書の関連文書から情報抽出を行う処理を、指定された最大リンク深度の範囲内で再帰的に行う抽出部とを備えたことを特徴とする情報抽出装置。

【0016】

〈構成 11〉

構成 5～7 のいずれかに記載の情報抽出装置において、最大リンク深度を指定する最大リンク深度指定部と、対象文書から情報抽出できなかった場合は、その文書の関連文書から情報抽出を行う処理を、指定された最大リンク深度の範囲内で再帰的に行う探索部とを備えたことを特徴とする情報抽出装置。

【0017】

〈構成 12〉

構成 10 に記載の情報抽出装置において、リンク深度の値が小さい文書から順に情報抽出処理を行う抽出部を備えたことを特徴とする情報抽出装置。

【0018】

〈構成 13〉

構成 11 に記載の情報抽出装置において、リンク深度の値が小さい文書から順に情報抽出処理を行う探索部を備えたことを特徴とする情報抽出装置。

【0019】

〈構成 14〉

構成 1～4、10、12 のいずれかに記載の情報抽出装置において、関連文書の文書アドレスに基づいて内部リンクと外部リンクとを判別し、外部リンクの文書は情報抽出の対象から除外する抽出部を備えたことを特徴とする情報抽出装置。

【0020】

〈構成 15〉

構成 5～7、11、13 のいずれかに記載の情報抽出装置において、関連文書の文書アドレスに基づいて内部リンクと外部リンクとを判別し、外部リンクの文

書は情報抽出の対象から除外する探索部を備えたことを特徴とする情報抽出装置。

#### 【0021】

##### 〈構成16〉

構成3または4に記載の情報抽出装置において、階層構造に基づいて、上位階層の抽出結果から下位階層の抽出結果の順番に複数の文字列を結合することにより処理結果の文字列を作成する処理部を備えたことを特徴とする情報抽出装置。

#### 【0022】

##### 〈構成17〉

構成7に記載の情報抽出装置において、階層構造に基づいて、上位階層の抽出結果から下位階層の抽出結果の順番に複数の文字列を結合することにより処理結果の文字列を作成する探索部を備えたことを特徴とする情報抽出装置。

#### 【0023】

##### 〈構成18〉

構成3、4、16に記載の情報抽出装置において、階層構造で表現された複数の文字列を合成する場合の所定の合成ルールを有し、合成ルールに従って処理結果の文字列を作成する処理部を備えたことを特徴とする情報抽出装置。

#### 【0024】

##### 〈構成19〉

構成7または17に記載の情報抽出装置において、階層構造で表現された複数の文字列を合成する場合の所定の合成ルールを有し、合成ルールに従って処理結果の文字列を作成する探索部を備えたことを特徴とする情報抽出装置。

#### 【0025】

#### 【発明の実施の形態】

以下、本発明の実施の形態を具体例を用いて詳細に説明する。

#### 《具体例1》

##### 〈構成〉

図1は、本発明の情報抽出装置の具体例1を示す構成図である。

図示の装置は、コンピュータで構成され、記憶部101、起点アドレス指定部

102、カテゴリ指定部103、最大リンク深度指定部104、バッファ部105、抽出部106、処理部107、リンク情報管理部108、表示部109を備えている。

#### 【0026】

記憶部101は、例えばハードディスク装置等の記憶装置からなり、処理対象の文書を記憶する機能部である。

#### 【0027】

図2は、記憶部101に記憶される文書の一例である。

図示例では、文書111～120までの20の文書を示しているが、実際にはその他の文書がもっと多く存在していても構わない。図中の矢印はリンクを表しており、矢印の元の文書が矢印の先の文書へのリンクを持っていることを示している。また、文書111～文書117は「xyz.jp」という同一サイト内部の文書である。尚、図中で、これらの文書のアドレスはサイト名を省略して記述している。例えば、文書111の文書アドレスは、一般的には「xyz.jp/A1.html」であるが、サイト名を省略して「A1.html」とだけ記述してある。文書118～文書120は「xyz.jp」というサイト以外の文書である。

#### 【0028】

図1に戻り、起点アドレス指定部102は、情報抽出を実施する対象文書のアドレスを利用者が指定する機能部である。カテゴリ指定部103は、利用者が抽出したい情報の種類（カテゴリ）を指定する機能部である。最大リンク深度指定部104は、利用者が情報抽出を実施する範囲を指定する機能部である。この範囲としては、例えば、リンク深度が2の場合は、起点文書のアドレスからリンクを2回参照してたどり着くことができる文書までが情報抽出を実施する範囲となる。尚、以上の起点アドレス指定部102～最大リンク深度指定部104は、例えば、キーボードやポインティングデバイス等の入力装置で構成されている。

#### 【0029】

バッファ部105は、抽出部106が抽出する場合や処理部107が処理を行うために、記憶部101から対象の1文書を取得し一時的に記憶する機能部であり、例えば主メモリ上の一領域で実現されている。

**【0030】**

抽出部106は、バッファ部105に記憶された文書からカテゴリ指定部103で指定された情報を抽出する機能部である。処理部107は、抽出部106に抽出の開始を指示し、抽出部106の抽出結果の有無に基づいて処理の流れを制御し、バッファ部105からリンク情報を取得してそれが内部サイトへのリンクであった場合はリンク情報管理部108に記録し、リンク情報管理部108のリンク情報に基づいて、次に処理すべき文書を記憶部101から取り出してバッファ部105にロードする機能部である。

**【0031】**

リンク情報管理部108は、リンク元文書のアドレスとリンク先文書のアドレスの関係を起点アドレスから始まるツリー構造で管理する機能部である。表示部109は、ディスプレイ等の表示装置とその制御部からなり、抽出部106が抽出した結果を表示するための機能部である。

**【0032】**

尚、上記の抽出部106～リンク情報管理部108は、それぞれの構成に対応したソフトウェアと、これらのソフトウェアを実行するためのCPUやメモリ等のハードウェアから実現されているものである。

**【0033】****〈動作〉**

図3は、具体例1の動作を示すフローチャートである。

以下、図のフローチャートに沿って動作を説明する。

まず、現在のリンク深度を表す変数であるリンク深度Dに0を代入する（ステップS101）。次に、起点アドレス指定部102で指定されたアドレスをリンク情報管理部108の先頭に設定する（ステップS102）。例えば、起点アドレス指定部102で「xyz.jp/A1.html」が起点アドレスに指定された場合、リンク情報管理部108のデータは次の通りである。

**【0034】**

図4は、リンク情報管理部108のデータの説明図（その1）である。

リンク情報管理部108は、サイト内部のリンクしか扱わないので、サイト名

部分は省略して表示している。次に、リンク情報管理部108のデータを参照しながら、リンク深度Dの全てのアドレスに対してステップS104からステップS108までの処理を繰り返す（ステップS103）。繰り返す内容は次の通りである。

#### 【0035】

先ず、処理部107は、バッファ部105にロードされた文書にリンクがあるかを調べて、文書中の全てのリンク先アドレスを取得し（ステップS105）、内部サイトへのリンクだけをリンク情報管理部108内の現在処理しているアドレスの下位アドレスとして設定する（ステップS106）。例えば、文書のリンク関係が図2の場合は、初めてステップS106を終了した時点で、リンク情報管理部108のデータは次のようになる。

#### 【0036】

図5は、リンク情報管理部108のデータの説明図（その2）である。

ここで、文書118は外部サイトへのリンクであるためリンク情報管理部108には設定されない。次に、抽出部106は、バッファ部105の文書からカテゴリ指定部103で指定されたカテゴリの情報を取得し、情報抽出を行う（ステップS107）。このステップS107において、抽出結果が得られた場合（ステップS108）は、これを表示部109で表示し（ステップS114）、処理を終了する。

#### 【0037】

一方、ステップS108において、抽出結果が得られなかった場合はステップS103に戻って、上述した処理を繰り返す（ステップS109）。ステップS103～ステップS109までの繰り返しが終了すると、処理部107は、リンク深度Dの値に1を加算し（ステップS110）、その結果が最大リンク深度指定部104で指定した値を超えていた場合（ステップS111）、または、ステップS111において指定した値を超えてはいないが、リンク情報管理部108内に次に処理すべきアドレスがない場合（ステップS112）は、抽出ができなかった旨の表示を行い（ステップS113）、処理を終了する。一方、ステップS112において、次に処理すべきアドレスがあった場合はステップS103に



戻って処理を繰り返す。

### 【0038】

例えば、文書のリンク関係が図2に示す場合で、最大リンク深度指定部104で指定するリンク深度Dが2で最後までカテゴリ指定部103で指定したカテゴリの情報が抽出できなかった場合、最終的にリンク情報管理部108のデータは次のようになる。

### 【0039】

図6は、リンク情報管理部108のデータの説明図（その3）である。

文書118～文書120は、それぞれ外部のサイトの文書アドレスなので、リンク情報管理部108には設定されない。尚、リンクの参照関係がループしているために、リンク情報管理部108のデータとして、文書118～文書113のアドレスが2回現れるが、処理上特に問題はない。

### 【0040】

#### 〈効果〉

以上のように、具体例1によれば、次のような効果がある。

●リンク先からも情報抽出を行うので、本来一つの文書にするはずのものを、読みやすさを向上させるために複数に分割し、それらを互いにリンクさせている場合であっても情報抽出を的確に実施することができる。

●リンク先が外部サイトの場合は情報抽出をしないように構成したので、参考のために指し示しているだけのリンクなどの場合はリンク先から情報を提出することがなく、本来一つの文書にする筈のものだけからの的確に情報抽出を行うことができる。

●最大リンク深度の指定により終了条件を設定するようにしたので、リンクの参照関係がループを構成している場合であっても問題なく動作する。

●リンク深度の値が小さい文書から順に情報抽出を行うようにしたので、より関連性の高い文書から処理することができ、抽出精度および処理速度を向上させることができる。これは、一般に、リンク深度の値が大きいほど対象文書と関連文書との関連性が下がっていく傾向があるためである。

●事前の処理が必要ないため、処理結果を保存しておく記憶容量を必要としな

い。また、要求のあった時点で処理を行うため、文書の最新の内容に対応することができる。

#### 【0041】

##### 《具体例2》

具体例2は、対象文書がディレクトリ構造で管理されている場合に、対象文書の一つ上のディレクトリにある特定の名前の文書を上位文書として、この上位文書も情報抽出の対象文書とするようにしたものである。

#### 【0042】

##### 〈構成〉

図7は、具体例2の構成図である。

図示の装置は、記憶部101、起点アドレス指定部102、カテゴリ指定部103、バッファ部105、抽出部106、表示部109、処理部201、カテゴリ階層規定部202からなる。ここで、処理部201およびカテゴリ階層規定部202以外の構成は具体例1と同様であるため、対応する部分に同一符号を付してその説明を省略する。

#### 【0043】

処理部201は、抽出部106に抽出の開始を指示し、抽出部106の抽出結果がカテゴリ階層の一部分のみの場合は、対象文書のアドレスから上位文書のアドレスを生成し、この上位文書から上位階層の情報を抽出することを繰り返し、最後に、これらの抽出結果をカテゴリ階層規定部202の階層構造の情報に基づいて合成して表示部109に出力する機能部である。また、カテゴリ階層規定部202は、抽出部106が参照するデータであり、抽出結果カテゴリの上下関係を階層構造で規定する機能部である。

#### 【0044】

尚、上記の処理部201は、それぞれの構成に対応したソフトウェアと、これらのソフトウェアを実行するためのCPUやメモリ等のハードウェアから実現されているものである。

#### 【0045】

##### 〈動作〉

図12は、具体例2の動作を示すフローチャートである。

以下、図のフローチャートに沿って動作を説明する。

先ず、処理部201により、起点アドレス指定部102が示す文書の内容をバッファ部105にロードする（ステップS201）。次に、抽出部106は、バッファ部105の文書からカテゴリ指定部103で指定されたカテゴリの情報を抽出する（ステップS202）。この抽出処理で抽出できなかった場合（ステップS203）は、その旨を表示して（ステップS204）、処理を終了する。また、抽出結果が完全な場合（一部分のみではない場合）は、抽出結果を表示して処理を終了する（ステップS205、ステップS206）。一方、ステップS205において、抽出結果が一部分のみであった場合、処理部201は、処理した文書のアドレスから上位の文書アドレスを生成し（ステップS207）、その文書が存在するかどうかを調べる（ステップS208）。

#### 【0046】

ステップS208において、文書が存在しない場合は、一部分のみの抽出結果を表示して（ステップS209）、処理を終了する。文書が存在する場合は、そのアドレスが示す文書の内容をバッファ部105にロードし（ステップS210）、バッファ部105の文書からカテゴリ指定部103で指定されたカテゴリで、かつ、ステップS202で抽出されたものよりも上位階層の情報を抽出する（ステップS211）。処理部201は、ステップS211の抽出処理において抽出できなかった場合（ステップS212）は、ステップS207に戻り、更にその文書の上位文書のアドレスを生成する。このように、ステップS212で情報が抽出できなかった場合はステップS207～ステップS212の処理を再帰的に繰り返す。また、ステップS212において、情報を抽出できた場合は以前の抽出結果と合成し（ステップS213）、その結果を表示して（ステップS214）、処理を終了する。

#### 【0047】

以下、一例を用いて更に詳細に動作を説明する。

図10は、ディレクトリ構造の説明図である。

図示のように、文書211～文書216を含む多くの文書が管理されていると

する。また、図10中の点線内部にある文書の参照関係は次のようになっている。

#### 【0048】

図8は、文書211～文書216の参照関係の説明図である。

図9は、文書211～文書216の内容を示す説明図である。

尚、図8では煩雑さを避けるため省略して記載しているが、実際にはディレクトリの名前なども文書アドレスに含まれる。例えば、文書211のアドレスを省略せずに示すと、「shousei.ac.jp/kgb/jhk/index.html」となる。

#### 【0049】

このような文書に対して、処理部201は、先ず、起点アドレス指定部102が示す文書の内容をバッファ部105にロードする（ステップS201）。今、起点アドレス指定部102がshousei.ac.jp/kgb/jhk/lab/02.htmlを示しているとする、抽出部106は、図9（c）に示すような内容をバッファ部105にロードする。

#### 【0050】

次に、抽出部106は、バッファ部105の文書からカテゴリ指定部103で指定されたカテゴリの情報を抽出する（ステップS202）。今、カテゴリとして「組織名」を指定しているとする、抽出部106は図9（c）の内容から組織名として「井上研究室」という単語を「研究室名」として抽出する。尚、この処理は、「…研究室」といった“研究室”を接尾語として含む文字列を抽出するといったことにより行うものである。次に、処理部201は、この結果をカテゴリ階層規定部202の組織名カテゴリの階層と比較する（ステップS203、S205）。

#### 【0051】

図11は、カテゴリ階層規定部202のデータの一例を示す説明図である。

図11を参照すると、「組織名」が完全であるためには、「大学名」「学部名」「研究室名」の四つの情報、または、「会社名」「部名」「課名」「係名」の四つの情報が揃っている必要があることが分かる。従って、この場合は「研究室名」しか抽出できなかったため、抽出結果は一部分のみであることになる。そこ

で、処理部 201 は、元の文書アドレスから上位文書のアドレスを生成する（ステップ S 206）。ここでは、上位文書は、一つ上のディレクトリの index.html という名前の文書であるとする。従って、元の文書アドレスは、shousei.ac.jp/kgb/jhk/lab/02.html だったので、その上位文書のアドレスは、shousei.ac.jp/kgb/jhk/index.html となる。従って、このアドレスが存在しているかを判定すると、この文書は文書 211 として存在しているため、上位文書として抽出する。

#### 【0052】

従って、処理部 201 は、図 9（a）に示すような内容をバッファ部 105 にロードし（ステップ S 210）、この文書から「研究室名」よりも上位階層の「組織名」を抽出する（ステップ S 211）。結果として「情報工学科」を「学科名」として抽出できたとすると、ステップ S 202 での抽出結果である「井上研究室」（研究室名）と、今抽出した「情報工学科」（学科名）をカテゴリ階層規定部 202 で示される順序で結合し、「情報工学科井上研究室」という単語を合成し（ステップ S 213）、それを表示して（ステップ S 214）、処理を終了する。

#### 【0053】

##### 〈効果〉

以上のように、具体例 2 によれば次のような効果が得られる。

●上位文書からも情報抽出を行うので、本来一つの文書にする筈のものを、読みやすさを向上させるために複数に分割し、それらを互いにリンクさせている場合であっても情報抽出を的確に実施することができる。

●リンクの情報は使わずに、ディレクトリ構造の情報だけを使うので、単純な処理で実現することができる。ディレクトリはツリー構造であり、リンクのようにループが構成されたりしないので、それらを解消するための処理を必要としない。

●二つの文書から抽出した単語を合成するので、文書中には存在しない単語を結果として出力することができる。更に、カテゴリ階層に基づいて合成するので、単語の合成を的確に実施することができる。

●事前の処理が必要ないので、処理結果を保存しておく記憶容量を必要としない。

い。また、文書の最新の内容に対応することができる。

#### 【0054】

##### 《具体例3》

具体例3は、具体例1と同等の結果を得るのに、文書収集時に情報抽出とリンク情報の取得を実施するように構成したものである。

#### 【0055】

##### 〈構成〉

図13は、具体例3の構成図である。

図の装置は、記憶部101、起点アドレス指定部102、カテゴリ指定部103、最大リンク深度指定部104、バッファ部105、抽出部106、表示部109、収集部301、登録部302、抽出結果記憶部303、探索部304を備えている。ここで、記憶部101～表示部109は、具体例1、2と同様の構成であるため、その説明は省略する。

#### 【0056】

収集部301は、記憶部101に新しく文書が登録された場合や、文書が変更された場合にこれを察知し、登録部302に登録させる機能部である。記憶部101がワールドワイドウェブ（WWW：インターネットを介して参照できる様々な文書）の場合は、一般にウェブロボットと呼ばれる文書収集装置と同等のものであってもよい。

#### 【0057】

登録部302は、収集部301が新しく収集した文書から抽出部106が情報抽出した結果とリンク先またはリンク元の情報を抽出結果記憶部303に登録する機能部である。例えば、図2のようなリンクで関連付けられた文書を登録した場合、抽出結果記憶部303内部のデータは次のようになる。

図14は、抽出結果記憶部303の内部データの説明図である。

但し、図14において、各文書の内容は例示していないので、抽出結果は仮に示したものである。

#### 【0058】

探索部304は、起点アドレス指定部102、カテゴリ指定部103、最大リ

ンク深度指定部 104 に設定された条件に基づいて抽出結果記憶部 303 から必要な情報を探索し、その結果を表示部 109 に出力する機能部である。

#### 【0059】

尚、上記の収集部 301、登録部 302 および探索部 304 は、それぞれの構成に対応したソフトウェアと、これらのソフトウェアを実行するための CPU やメモリ等のハードウェアから実現されているものである。

#### 【0060】

##### 〈動作〉

具体例 3 の動作として、登録時の動作と探索時の動作それぞれについて順に説明する。

図 16 は、具体例 3 における登録時の動作を示すフローチャートである。

収集部 301 が処理対象の文書を発見すると、まず、対象文書をバッファ部 105 にロードする（ステップ S301）。次に、抽出部 106 が情報抽出を実施する（ステップ S302）。このとき、カテゴリ指定部 103 の内容にかかわらず、全てのカテゴリに対して抽出を行う。更に、登録部 302 はリンク先およびリンク元の情報を取得し（ステップ S303）、ステップ S302 で得た情報抽出の結果と共に抽出結果記憶部 303 に記憶させて（ステップ S304）、処理を終了する。その処理結果が図 14 に示す状態である。以上の動作を収集部 301 が処理対象の文書を発見する度に実施する。

#### 【0061】

図 17 は、具体例 3 の探索時の動作を示すフローチャートである。

まず、探索部 304 において、現在のリンク深度を表す変数であるリンク深度 D に 0 を代入する（ステップ S311）。次に、リンク深度 D の値に基づいて対象文書リストを作成する（ステップ S312）。対象文書リストとは、起点アドレス指定部 102 からリンク深度 D の回数だけリンク先またはリンク元をたどっていける文書のリストのことである。例えば、文書のリンク関係が図 2 のようになっているときに、起点アドレス指定部 102 により、起点アドレスに xyz.jp/A3.html が指定された場合、各リンク深度 D の対象文書リストは次のようになる。

#### 【0062】

図15は、対象文書リストの説明図である。

尚、具体例3でも具体例1と同様に外部サイトへのリンクは対象としないようにする。

#### 【0063】

次に、対象文書に、探索部304は、カテゴリ指定部103で指定されたカテゴリの抽出結果が存在するかどうか抽出結果記憶部303を参照して調べ（ステップS313）、あった場合はその結果を表示して（ステップS318）、処理を終了する。なかった場合は、リンク深度Dの値に1を加算し（ステップS315）、その結果が最大リンク深度指定部104の示す値を超えていた場合は、抽出できなかった旨を表示し（ステップS317）、処理を終了する。そうでない場合は、ステップS313へ戻って処理を繰り返す。

#### 【0064】

〈効果〉

以上のように、具体例3によれば、次のような効果が得られる。

●リンク先からも情報抽出を行うので、本来一つの文書にする筈のものを読みやすさを向上させるために複数に分割し、それらを互いにリンクさせている場合であっても、情報抽出を的確に実施することができる。

●リンク先が外部サイトの場合は、情報抽出をしないように構成してあるので、参考のために指し示しているだけのリンクなどの場合は、リンク先から情報を抽出することがなく、本来一つの文書にする筈のものだけからの的確に情報抽出することができる。

●最大リンク深度の指定により終了条件が設定されるので、リンクの参照関係がループを構成している場合であっても問題なく動作する。

●リンク深度の値が小さい文書から順に情報抽出を行うようにしたので、より関連性の高い文書から処理することができ、抽出精度および処理速度を向上させることができる。

●事前にリンク先の文書アドレスを収集しているので、全ての文書の事前処理が終了すれば、リンク元の文書アドレスの情報も完全に収集することができる。このため、参照元の文書からの情報抽出結果も利用することができる。



●事前の情報抽出の処理を完了しているので、応答が速い。

#### 【0065】

##### 《具体例4》

具体例4は、具体例2と同等の結果を得るのに、文書収集時に情報抽出とリンク情報および上位文書アドレスの取得を実施するようにしたものである。更に、上位文書には具体例2で説明した一つ上のディレクトリに存在する特定の名前の文書以外に、リンク元の文書が一つ上のディレクトリにある場合にはその文書を上位文書とするよう構成した。

#### 【0066】

##### 〈構成〉

図18は、具体例4の構成図である。

図の装置は、記憶部101、起点アドレス指定部102、カテゴリ指定部103、バッファ部105、抽出部106、表示部109、カテゴリ階層規定部202、収集部301、登録部401、抽出結果記憶部402、探索部403を備えている。ここで、記憶部101～表示部109は、具体例1の構成と同様であり、また、カテゴリ階層規定部202は具体例2、収集部301は具体例3の構成と同様であるため、ここでの説明は省略する。

#### 【0067】

登録部401は、収集部301が新しく収集した文書から抽出部106が情報抽出した結果と、文書の内容から取得したリンク先またはリンク元の情報と、生成した上位文書の文書アドレスを抽出結果記憶部402に記憶する機能部である。抽出結果記憶部402は、各文書の抽出結果とリンク先またはリンク元の文書アドレスの情報と上位文書の文書アドレスを管理する機能部である。例えば、図8のようにリンクで関連付けられた文書を登録した場合、抽出結果記憶部402内部のデータは次のようになる。

#### 【0068】

図19は、抽出結果記憶部402内部のデータの説明図である。

但し、具体例4においても、図8と同様に文書アドレスの上位のディレクトリ名などは省略して示している。

**【0069】**

探索部403は、起点アドレス指定部102、カテゴリ指定部103に設定された条件に基づいて抽出結果記憶部402から必要な情報を探索すると共に、必要があれば探索の結果得られた抽出結果の単語をカテゴリ階層規定部202の階層に基づいて合成し、その結果を表示部109に出力する機能部である。

**【0070】**

尚、上記の登録部401および探索部403は、それぞれの構成に対応したソフトウェアと、これらのソフトウェアを実行するためのCPUやメモリ等のハードウェアから実現されているものである。

**【0071】****〈動作〉**

具体例4の動作として、登録時の動作と探索時の動作それぞれについて順に説明する。

図20は、具体例4における登録時の動作を示すフローチャートである。

収集部301が処理対象の文書を発見すると、まず、対象文書をバッファ部105にロードする（ステップS401）。次に、抽出部106が情報抽出を実施する（ステップS402）。このとき、カテゴリ指定部103の内容にかかわらず、全てのカテゴリに対して抽出を行う。次に、登録部401は、リンク先およびリンク元の情報を取得し（ステップS403）、更に、上位文書アドレスを生成する（ステップS404）。尚、上位文書には、具体例2で説明した一つ上のディレクトリに存在する特定の名前の文書以外に、リンク元の文書が一つ上のディレクトリにある場合にはその文書も上位文書とする。つまり、具体例2では上位文書の個数は最大でも一つであったが、具体例4では複数になる場合がある。

**【0072】**

最後に、ステップS402で得た情報抽出の結果と、ステップS403で得たリンク先およびリンク元の情報と、ステップS404で得た上位文書アドレスを抽出結果記憶部402に記憶させて（ステップS405）、処理を終了する。図19が、処理終了後の抽出結果記憶部402の内部データを示している。以上の動作を収集部301が処理対象の文書を発見する度に実施する。

## 【0073】

図21は、具体例4の探索時の動作を示すフローチャートである。

先ず、探索部403は、起点アドレス指定部102が示す文書からカテゴリ指定部103で指定されたカテゴリ情報の抽出結果が抽出結果記憶部402に存在するかどうかを探索する（ステップS411）。存在しなかった場合は、抽出できなかった旨を表示部109によって表示し（ステップS413）、処理を終了する。また、存在した抽出結果が完全な場合（一部分のみではない場合）は、抽出結果を表示して処理を終了する（ステップS415）。

## 【0074】

一方、抽出結果が一部分のみの場合は、抽出結果記憶部402の該当部分に登録された全ての上位文書アドレスに対して（ステップS416）、カテゴリ指定部103で指定されたカテゴリで、かつ、ステップS411で取得したものよりも上位階層の抽出結果が抽出結果記憶部402に存在するかどうかを探索する（ステップS417）。この探索で、存在した場合（ステップS418）は、以前に取得した抽出結果と合成し（ステップS419）、その結果を表示して（ステップS420）、処理を終了する。ステップS418において、存在しない場合はステップS417、S418を繰り返し（ステップS421）、繰り返しが終了した場合は、一部分のみの抽出結果を表示して（ステップS422）、処理を終了する。

## 【0075】

以下、一例を用いて探索時の動作を更に詳細に説明する。

この例では、記憶部101内部において、図10のようなディレクトリ構造で文書211～文書216を含む多くの文書が管理されているとする。また、図10の点線内部にある文書の参照関係は図8に示す通りであるとする。尚、図8では煩雑さを避けるため省略して記載しているが、実際にはディレクトリの名前なども文書アドレスに含まれる。例えば、文書211のアドレスを省略せずに示すと、「shousei.ac.jp/kgb/jhk/index.html」となる。登録時の動作を実行すると抽出結果記憶部402の内容は図19のようになっている。

## 【0076】

起点アドレス指定部 102 が、shousei.ac.jp/kgb/jhk/lab/02.html を指定し、また、カテゴリ指定部 103 がカテゴリとして「組織名」を指定していると、探索部 403 は、抽出結果記憶部 402 における 5 行目の抽出結果の列を参照し、組織名として「井上研究室」という単語を「研究室名」として抽出した結果を取得する（ステップ S411）。これをカテゴリ階層規定部 202 の「組織名」カテゴリの階層と比較する（ステップ S414）。カテゴリ階層規定部 202 のデータは図 11 に示す通りである。

#### 【0077】

この図 11 を参照すると、「組織名」が完全であるためには、「大学名」「学部名」「学科名」「研究室名」の四つの情報、または、「会社名」「部名」「課名」「係名」の四つの情報が揃っている必要があることが分かる。従って「研究室名」しか抽出できなかったので、抽出結果は一部分のみであることになり、ステップ S416 に進む。次に、探索部 403 は、抽出結果記憶部 402 における 5 行目の上位文書の列を参照することにより、上位文書は、shousei.ac.jp/kgb/jhk/shokai.html および shousei.ac.jp/kgb/jhk/index.html であることを知る。これらに対して探索部 403 は探索処理を実施する（ステップ S416）。

#### 【0078】

先ず、shousei.ac.jp/kgb/jhk/shokai.html を対象とすると、抽出結果記憶部 402 の 2 行目を参照することにより、組織名として「秋山研究室」「井上研究室」「遠藤研究室」という三つの単語を「研究室名」として抽出した結果を得ることができるが、これらはいずれもステップ S411 で得た「研究室名」よりも上位階層ではないので、必要な単語を取得できなかったとしてステップ S421 へ進み、次の shousei.ac.jp/kgb/jhk/index.html を対象とする。同様に、抽出結果記憶部 402 の 1 行目も参照することにより、組織名として「情報工学科」という単語を「学科名」として抽出した結果を得ることができる。これは、カテゴリ階層規定部 202 を参照することにより、ステップ S411 で得た「研究室名」の上位階層にあたることが分かるので、対象とする単語が存在したとしてステップ S419 へ進む。

#### 【0079】

ステップ S 4 1 1 で得た「井上研究室」（研究室名）と、ステップ S 4 1 7 で得た「情報工学科」（学科名）をカテゴリ階層規定部 2 0 2 で示される順序で結合し、「情報工学科井上研究室」という単語を合成し（ステップ S 4 1 9）、それを表示して（ステップ S 4 2 0）、処理を終了する。

#### 【0080】

##### 〈効果〉

以上のように、具体例 4 によれば、次のような効果がある。

●上位文書からも情報抽出を行うので、本来一つの文書にする筈のものを、読みやすさを向上させるために複数に分割し、それらを互いにリンクさせている場合であっても情報抽出を的確に行うことができる。

●ディレクトリ構造の情報とリンクの参照元の情報とを組み合わせる使うので、リンク情報だけのときのようにループが構成されたりしないので、それらを解消するための処理を必要としない。

●二つの文書から抽出した単語を合成するので、文書中に存在しない単語を結果として出力することができる。更にカテゴリ階層に基づいて合成するので、単語の合成を的確に実施することができる。

●事前にリンク先の文書アドレスを収集しているので、全ての文書の事前処理が終了すれば、リンク元の文書アドレスの情報も完全に収集することができる。このため、参照元の文書からの情報抽出結果も利用することができる。

●事前の情報抽出の処理を完了しているので、応答が速い。

#### 【0081】

##### 《利用形態》

◆具体例 3 および具体例 4 では理解を助けるために、抽出結果記憶部 3 0 3、4 0 2 のデータとして、リンク元文書の文書アドレスを記憶する項目を設けて説明したが、この項目は必須ではない。抽出結果記憶部 3 0 3（4 0 2）に、リンク先文書のアドレスを記憶する項目さえあれば、これから逆にリンク元文書のアドレスを探すことは容易に可能である。

#### 【0082】

◆具体例 4 では、理解を助けるため、抽出結果記憶部 4 0 2 のデータ構造とし

て上位文書を記憶する項目を設けて説明したが、この項目は必ずしも必要な訳ではない。具体例2のように、必要になった時点で生成するようにしてもよい。

#### 【0083】

◆具体例2において、説明を分かり易くするため、上位文書から上位階層の情報を抽出できれば抽出処理を終了するよう説明した。つまり、単語を合成する数は最大でも二つという説明であったが、上位階層の情報を抽出できた後も、更に上位の階層の情報を対象文書の上位文書から抽出することを続けて、抽出できた全ての単語を合成するようにしてもよい。つまり、三つ以上の単語を合成する場合があってもよい。

#### 【0084】

◆具体例4において、説明を簡略化するため、上位文書を対象文書とすることを再帰的に繰り返すことは説明しなかったが、具体例2のステップS207～ステップS212の処理と同様に再帰的に繰り返すようにしてもよい。また、上述したように上位階層の情報を取得できた後も繰り返して取得して、三つ以上の単語を合成するようにしてもよい。

#### 【0085】

◆具体例4において、上位文書は対象文書の一つ上のディレクトリに存在する特定の名前の文書と、対象文書のリンク元の文書で、かつ、対象文書の一つ上のディレクトリに存在する文書の両方であると説明したが、これらのうちの片方だけを上位文書としてもよい。

#### 【0086】

◆具体例1～4において、記憶部101は、WWW（ワールドワイドウェブ）といったネットワーク上の文書であってもよいし、ハードディスク装置等の記憶装置内に格納された文書等、文書が存在する場所であれば、どのような形態であってもよい。

#### 【0087】

◆具体例1では、リンク先の文書から情報を抽出すると説明したが、これに限定されるものではない。これ以外にも具体例2や具体例4で説明した上位文書を対象にしてもよいし、リンク先の文書と上位文書の両方を対象としてもよい。

## 【0088】

◆具体例3では、リンク先の文書とリンク元の文書の両方から情報抽出結果を取得すると説明したが、具体例2や具体例4で説明した上位文書を対象に加えてもよい。更に、リンク先の文書、リンク元の文書、上位文書の3種類の文書から選んだ一つの文書または二つ以上の文書の組み合わせを対象としてもよい。

## 【0089】

◆具体例2や具体例4において、起点文書から抽出した単語と上位文書から抽出した単語を合成するように説明したが、これに限定されるものではない。同一文書から抽出された単語を合成してもよいし、具体例1や具体例3で説明したような、リンク先の文書やリンク元の文書から抽出した単語を合成してもよい。

## 【0090】

◆具体例2や具体例4において、抽出結果を合成する場合にカテゴリ階層規定部202の記載順序に従って単語を連結するよう説明したが、抽出した単語を連結する順序を別途合成ルールとして定義するよう構成してもよい。この合成ルールとは、連結順序を特定するものであればどのようなものであってもよいが、例えば次のような合成ルールである。

## 【0091】

例えば、情報としての地名が以下のように抽出できたとする。

<都道府県名>=大阪府

<市名>=大阪市

<区名>=浪速区

<町名>=日本橋

## 【0092】

ルールA

<都道府県名>+<市名>+<区名>+<町名>

ルールB

<町名>+“(“+<都道府県名>+”)”

という二つのルールがあった場合、

## 【0093】

ルール A の処理結果：大阪府大阪市浪速区日本橋

ルール B の処理結果：日本橋（大阪府）

といった結果となる。

#### 【0094】

ここでは、正確な住所を表記したい場合はルール A が、簡単に町名を特定して表記したい場合はルール B が有効である。

#### 【0095】

◆具体例 2 や具体例 4 において、上位文書として、一般的に上位文書であるとして用いられている index.html としたが、これに限定されるものではなく、予め、特定の名前の文書を決定するものであれば、どのような文書としてもよい。

#### 【0096】

◆具体例 1 ～ 4 において、表示部 109 は、ディスプレイ等の表示装置で表示を行う機能部であるとしたが、例えば印刷装置で印刷出力を行う機能部であってもよい。

◆具体例 1 ～ 具体例 4 をそれぞれ二つ乃至四つを任意に組み合わせてもよい。

#### 【0097】

##### 【発明の効果】

以上のように、本発明によれば、ハイパーテキスト構造を持つ文書群から指定情報を抽出する場合、ある起点アドレスの文書から情報が抽出できなかった場合は、その文書の関連文書から情報抽出を行うようにしたので、例えば、本来一つの文書にする筈のものを複数に分割し、それらを互いにリンクさせているような場合であっても情報抽出を的確に実施することができる。

##### 【図面の簡単な説明】

##### 【図 1】

本発明の情報抽出装置の具体例 1 を示す構成図である。

##### 【図 2】

記憶部に記憶される文書の一例を示す説明図である。

##### 【図 3】

具体例 1 の動作を示すフローチャートである。



**【図 4】**

リンク情報管理部のデータの説明図（その 1）である。

**【図 5】**

リンク情報管理部のデータの説明図（その 2）である。

**【図 6】**

リンク情報管理部のデータの説明図（その 3）である。

**【図 7】**

具体例 2 の構成図である。

**【図 8】**

文書 211～文書 216 の参照関係の説明図である。

**【図 9】**

文書 211～文書 216 の内容を示す説明図である。

**【図 10】**

ディレクトリ構造の説明図である。

**【図 11】**

カテゴリ階層規定部のデータの一例を示す説明図である。

**【図 12】**

具体例 2 の動作を示すフローチャートである。

**【図 13】**

具体例 3 の構成図である。

**【図 14】**

具体例 3 の抽出結果記憶部の内部データの説明図である。

**【図 15】**

対象文書リストの説明図である。

**【図 16】**

具体例 3 における登録時の動作を示すフローチャートである。

**【図 17】**

具体例 3 の探索時の動作を示すフローチャートである。

**【図 18】**

具体例 4 の構成図である。

【図 1 9】

具体例 4 の抽出結果記憶部の内部データの説明図である。

【図 2 0】

具体例 4 における登録時の動作を示すフローチャートである。

【図 2 1】

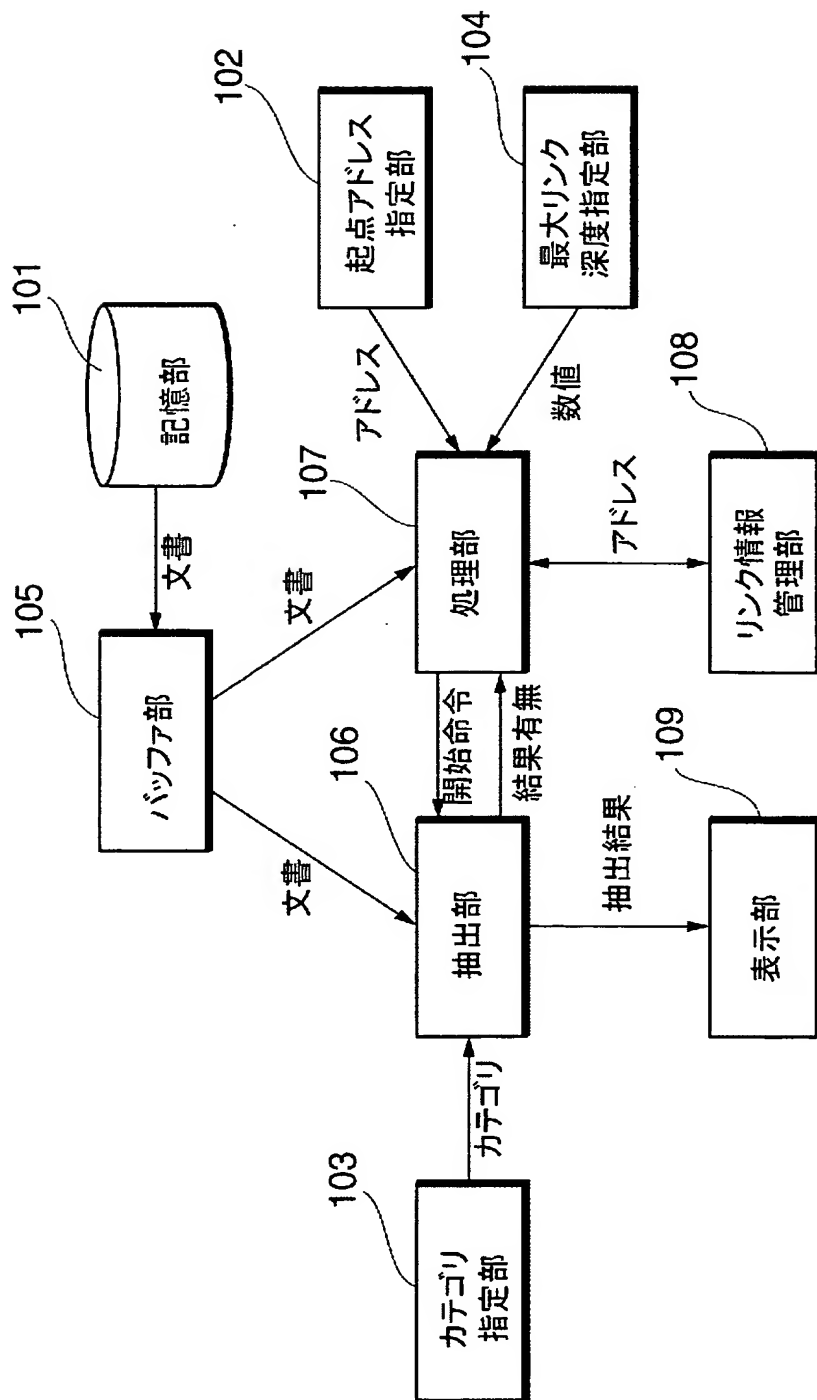
具体例 4 の探索時の動作を示すフローチャートである。

【符号の説明】

- 1 0 2 起点アドレス指定部
- 1 0 3 カテゴリ指定部
- 1 0 4 最大リンク深度指定部
- 1 0 6 抽出部
- 1 0 7、2 0 1 処理部
- 2 0 2 カテゴリ階層規定部
- 3 0 3、4 0 2 抽出結果記憶部
- 3 0 4 探索部

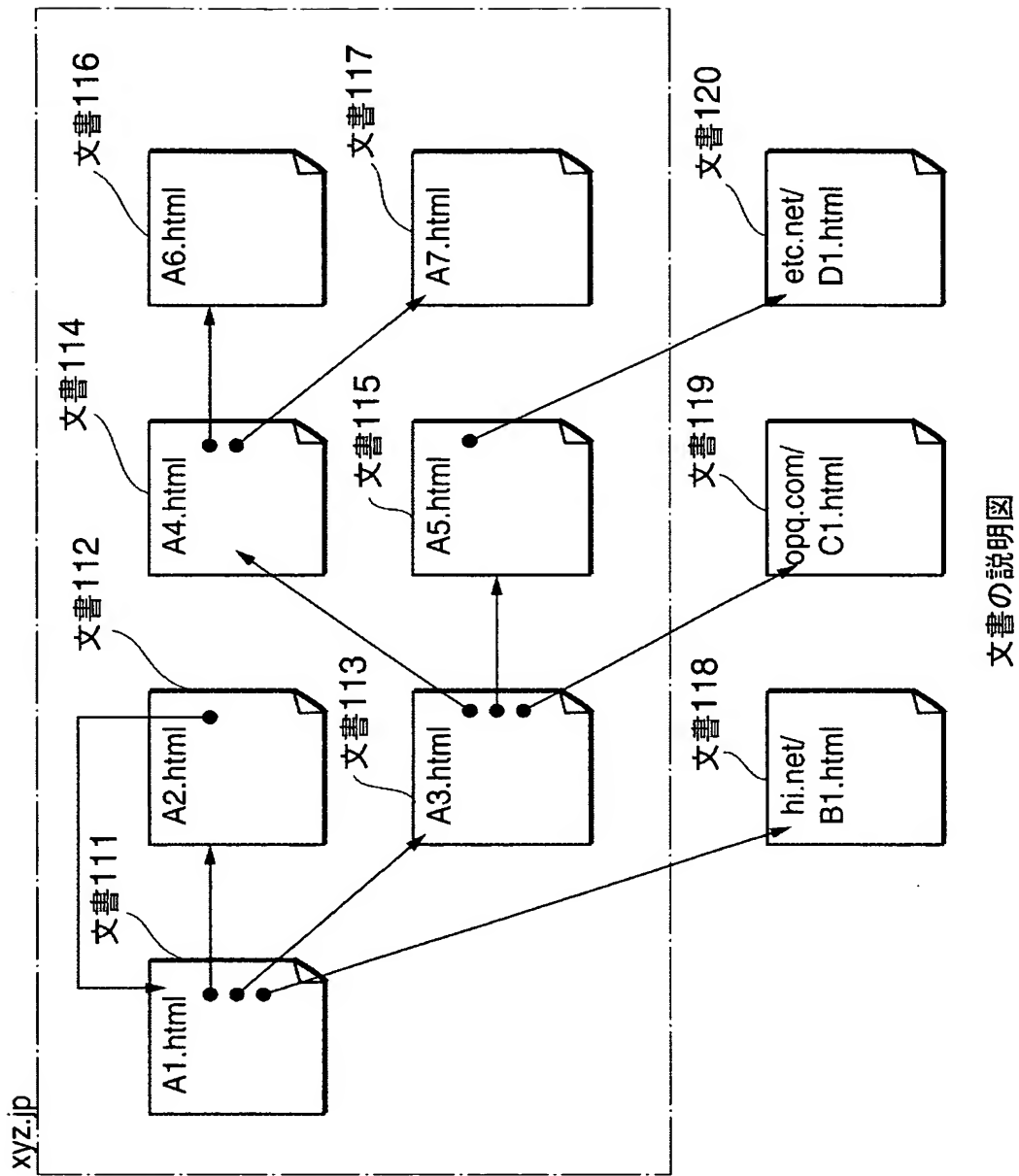
【書類名】 図面

【図 1】

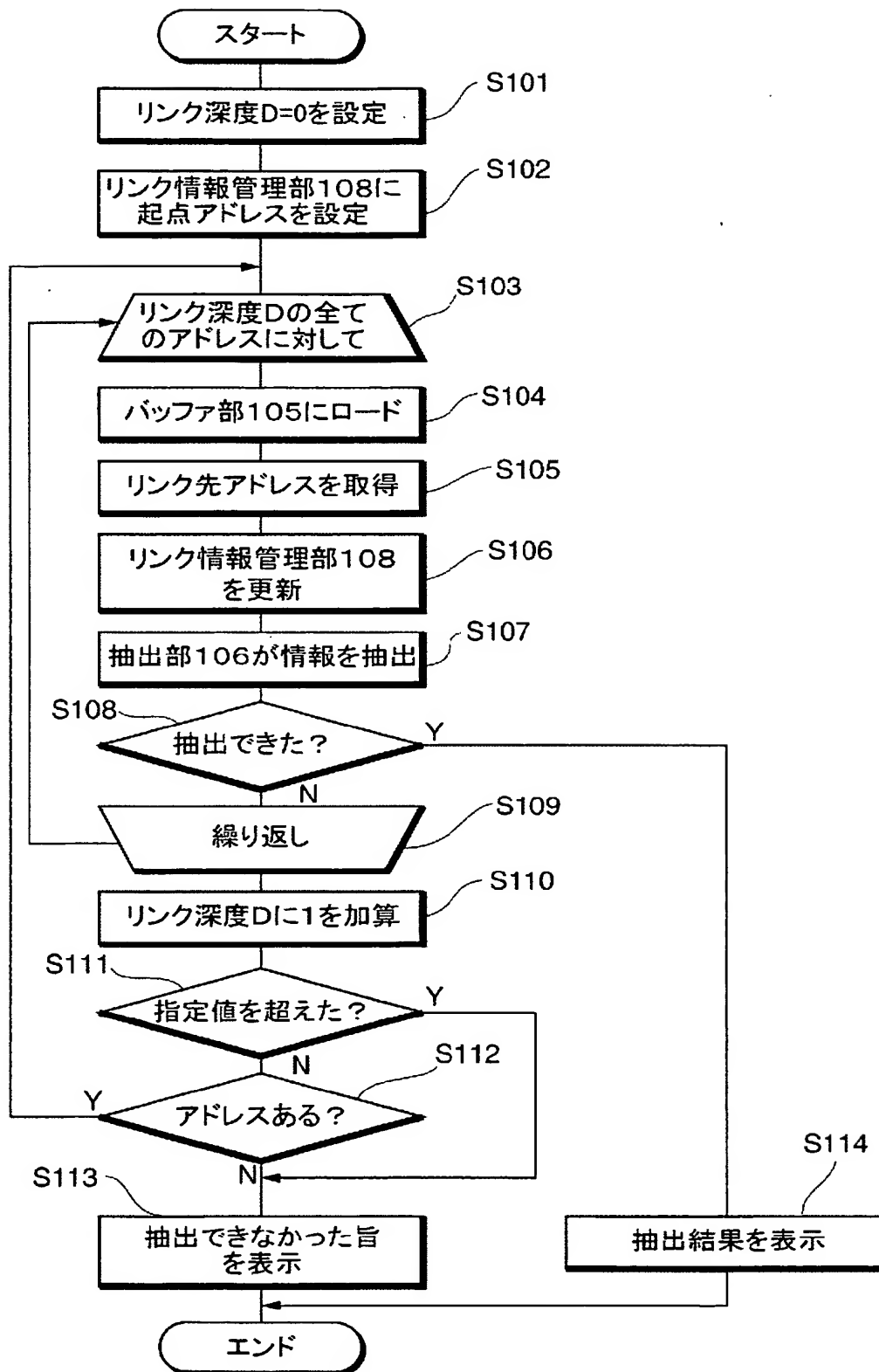


具体例1の構成図

【図 2】



【図 3】



具体例1の動作フローチャート

【図 4】

リンク深度 (D)	0	1	2	3
	A1.html			

リンク情報管理部のデータの説明図(その1)

【図 5】

リンク深度 (D)	0	1	2	3
	A1.html	A2.html A3.html		

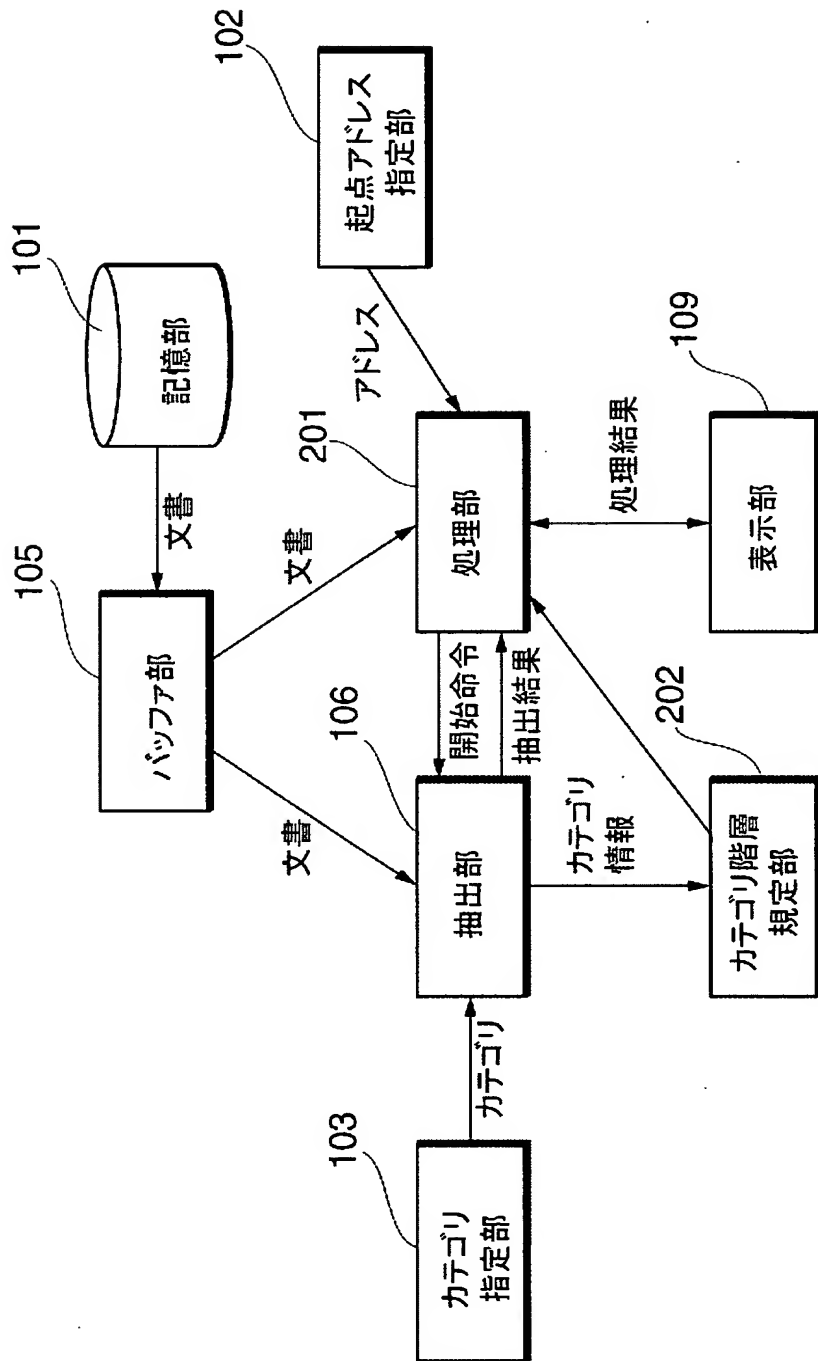
リンク情報管理部のデータの説明図(その2)

【図 6】

リンク深度 (D)	0	1	2	3
	A1.html	A2.html A3.html	A1.html A4.html A5.html	A2.html A3.html A6.html A7.html

リンク情報管理部のデータの説明図(その3)

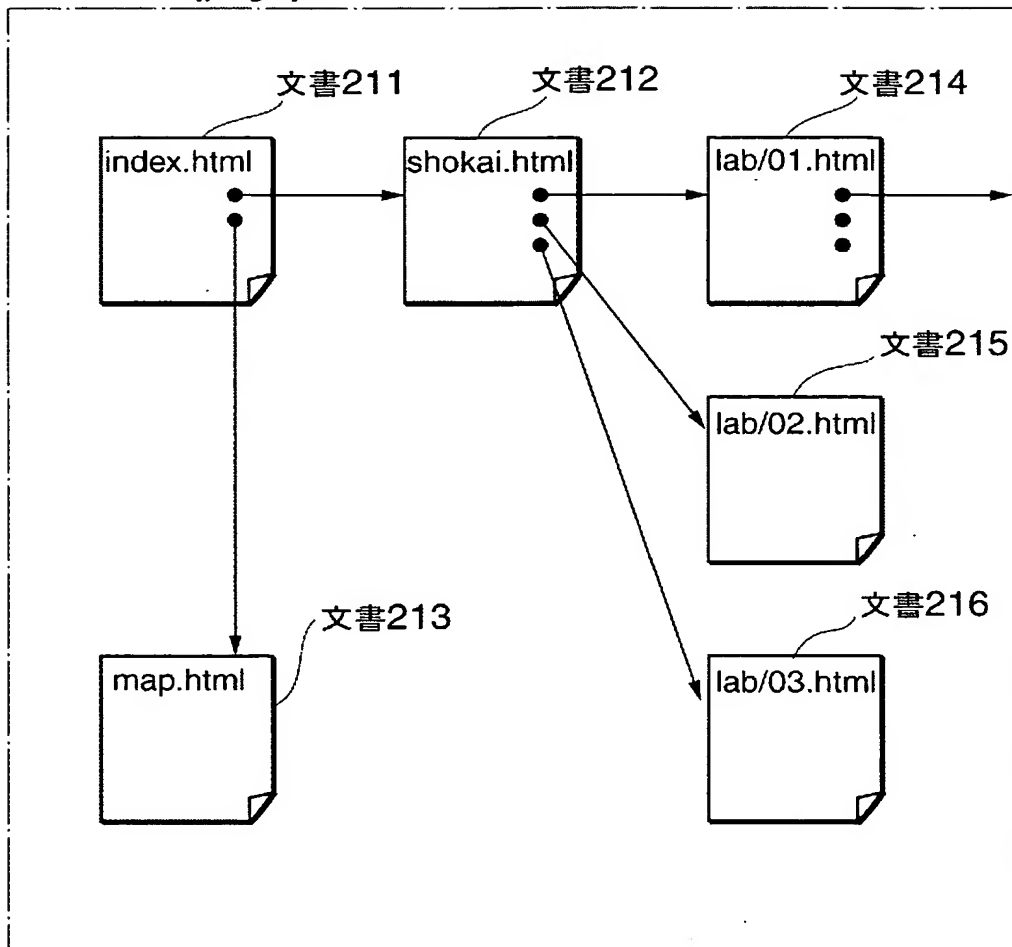
【図 7】



具体例2の構成図

【図 8】

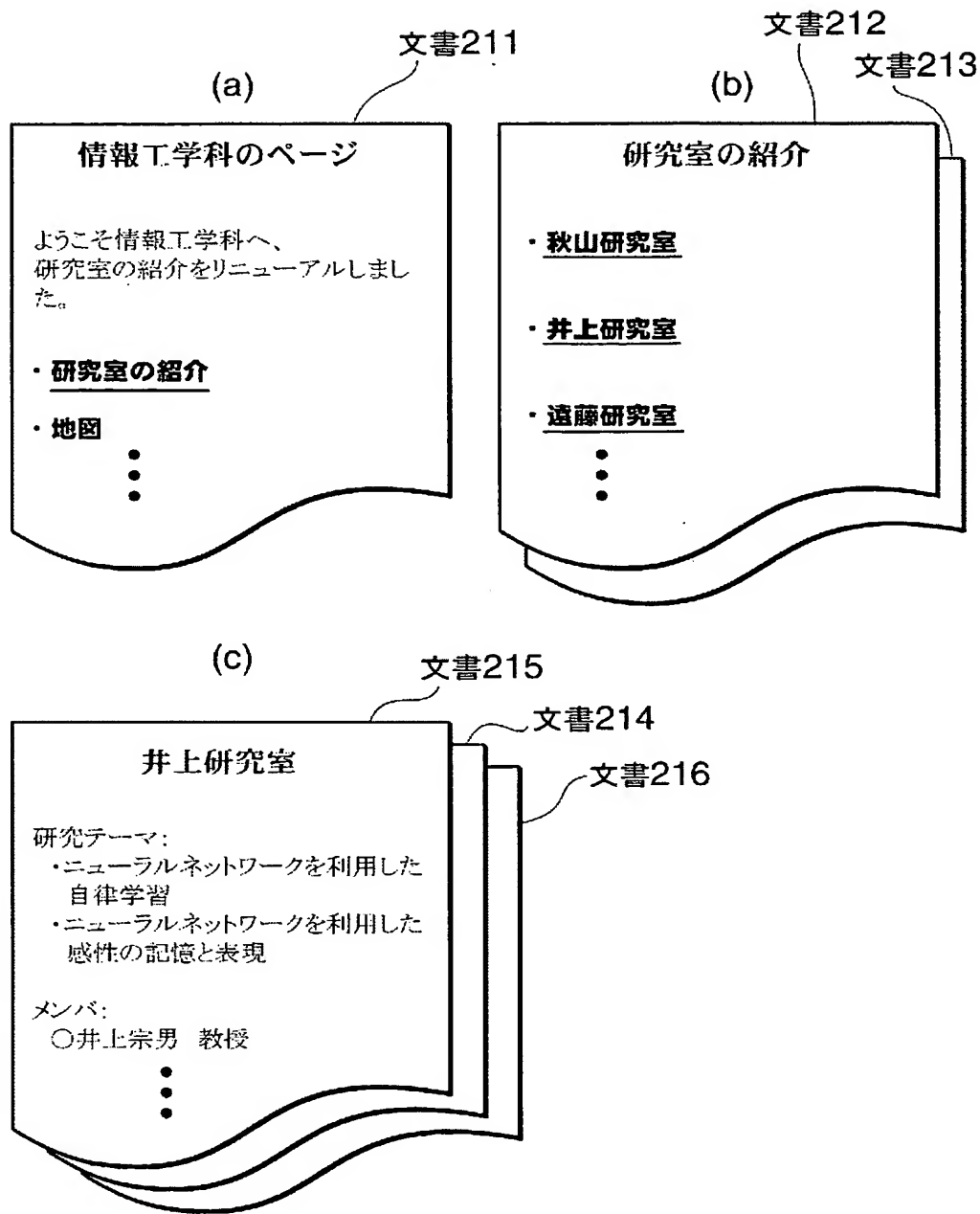
shousei.ac.jp/kgb/jhk



文書211～文書216の参照関係の説明図

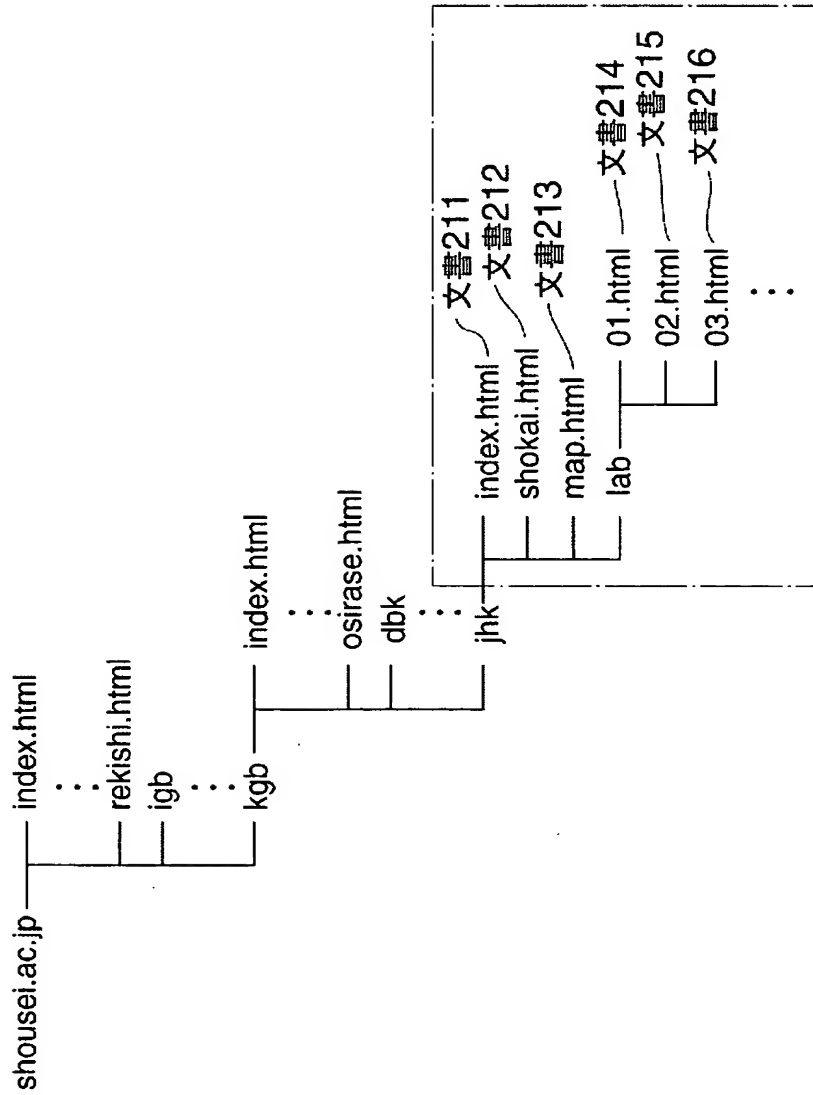


【図 9】



文書211～文書216の内容を示す説明図

【図 10】



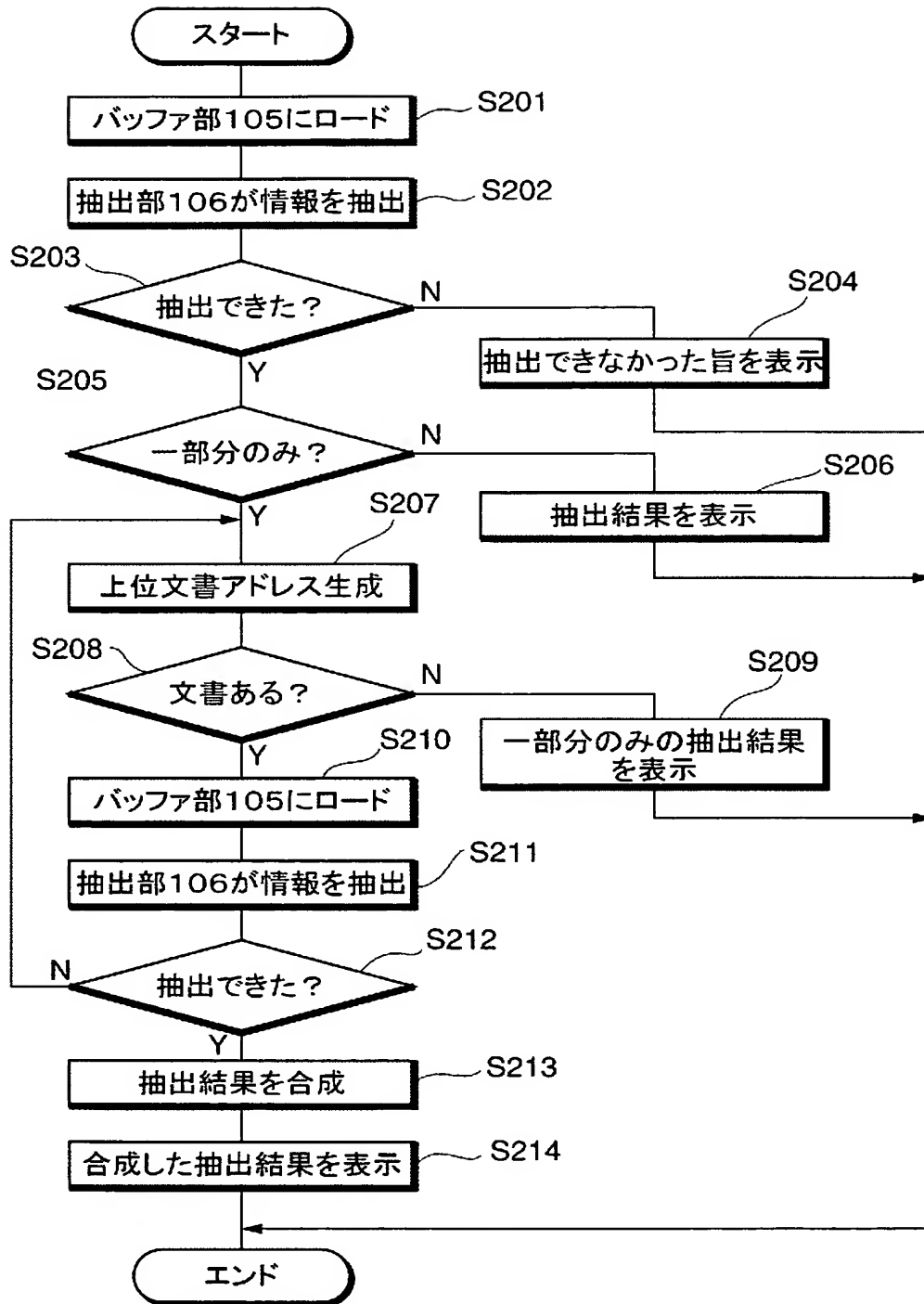
ディレクトリ構造の説明図

【図 11】

	上位階層←		←下位階層	
	大学名	学部名	学科名	研究室名
組織名(大学) (会社)	会社名	部名	課名	係名
	都道府県名	市名	区名	町名
住所	都道府県名	郡名	町名	—
	都道府県名	郡名	村名	—
駅名	鉄道会社名	線名	駅名	—
	属名	科名	目名	—
動植物の分類				

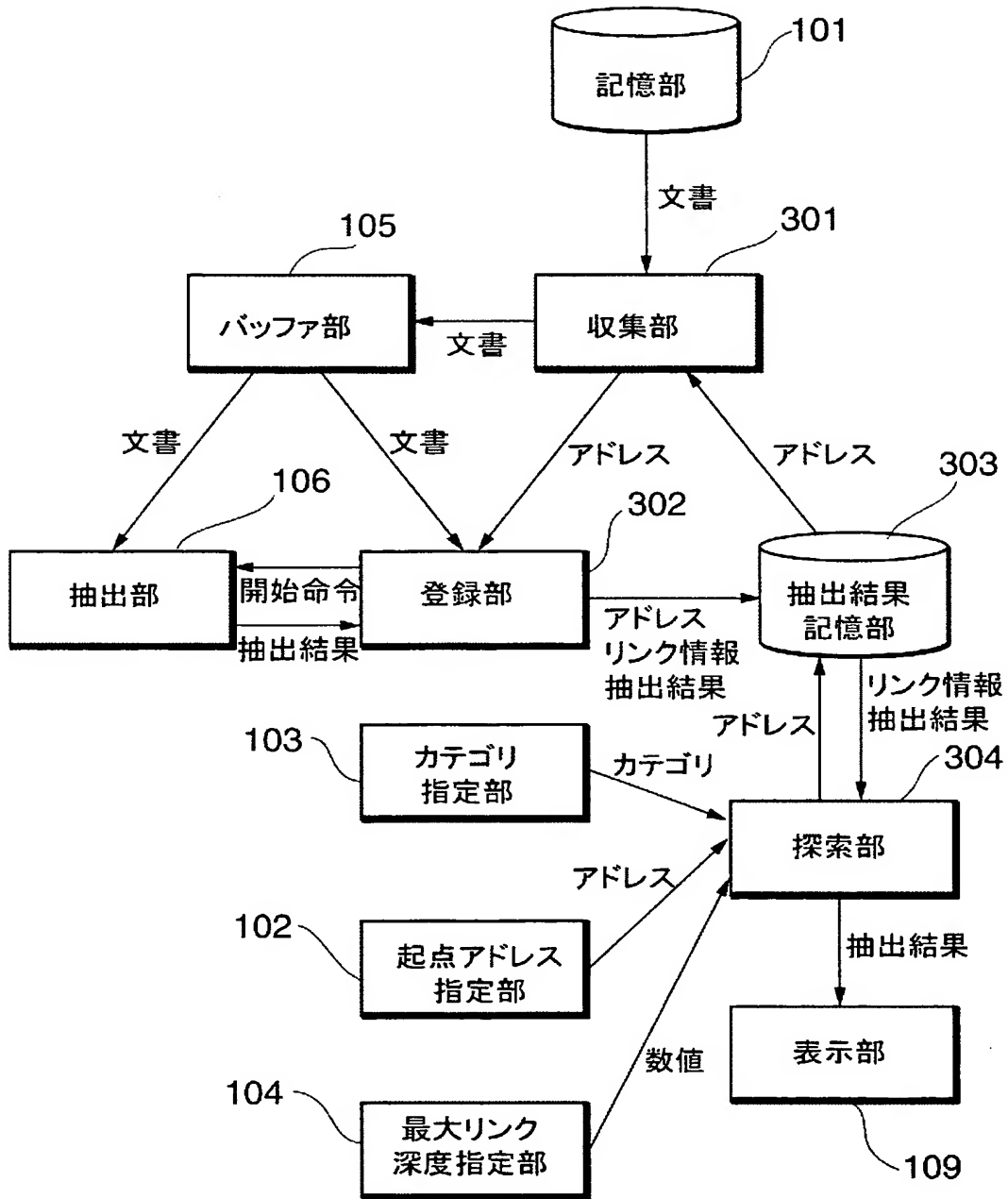
カテゴリ階層規定部のデータの説明図

【図 12】



具体例2の動作フローチャート

【図13】



具体例3の構成図

【図 14】

	アドレス	抽出結果		リンク情報	
		カテゴリ	抽出結果	リンク先	リンク元
1	A1.html	人名	植田稲男	A2.html A3.html	A2.html
2	A2.html	日付 日付	12/20 10/3	A1.html	A1.html
3	A3.html			A4.html A5.html	A1.html
4	A4.html	組織名(大学 名)	昭成大学	A6.html A7.html	A3.html
5	A5.html				A3.html
6	A6.html	人名	遠藤豆太郎		A4.html
7	A7.html				A4.html

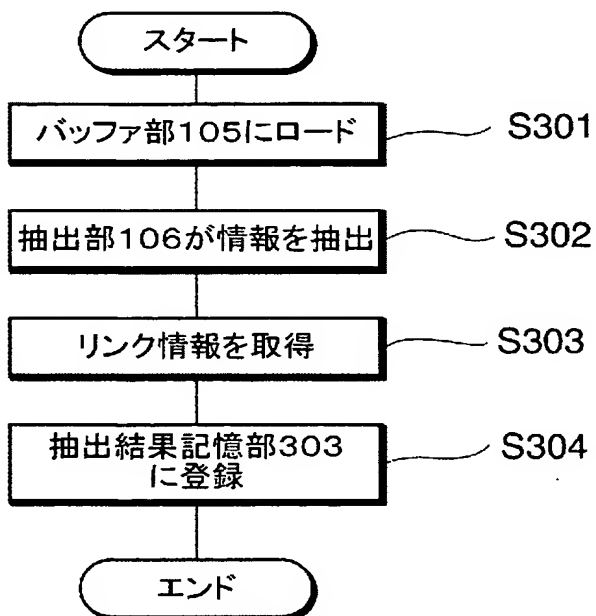
具体例3の抽出結果記憶部の内部データの説明図

【図15】

リンク深度D	文書リスト
0	A3.html
1	A4.html,A5.html,A1.html
2	A6.html,A7.html,A2.html

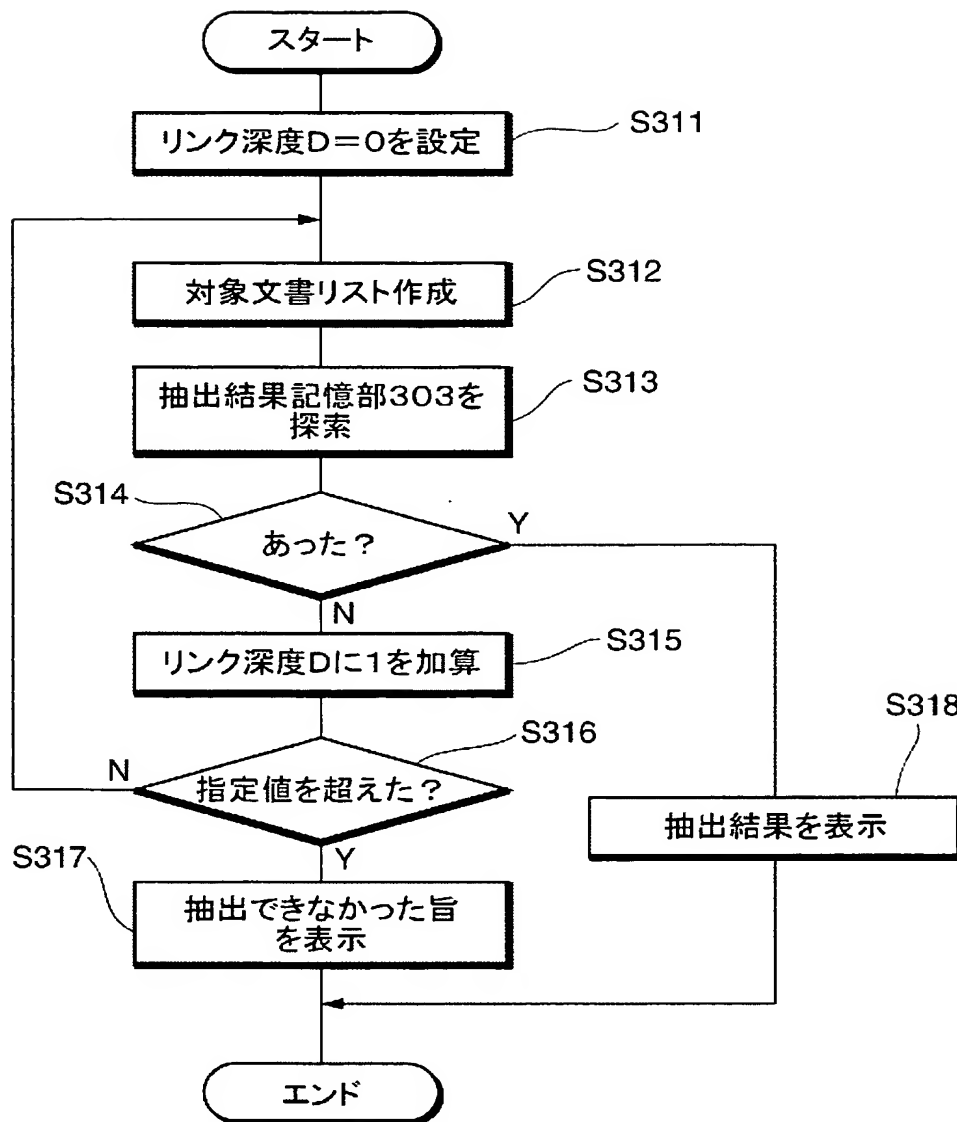
対象文書リストの説明図

【図16】



具体例3の登録時のフローチャート

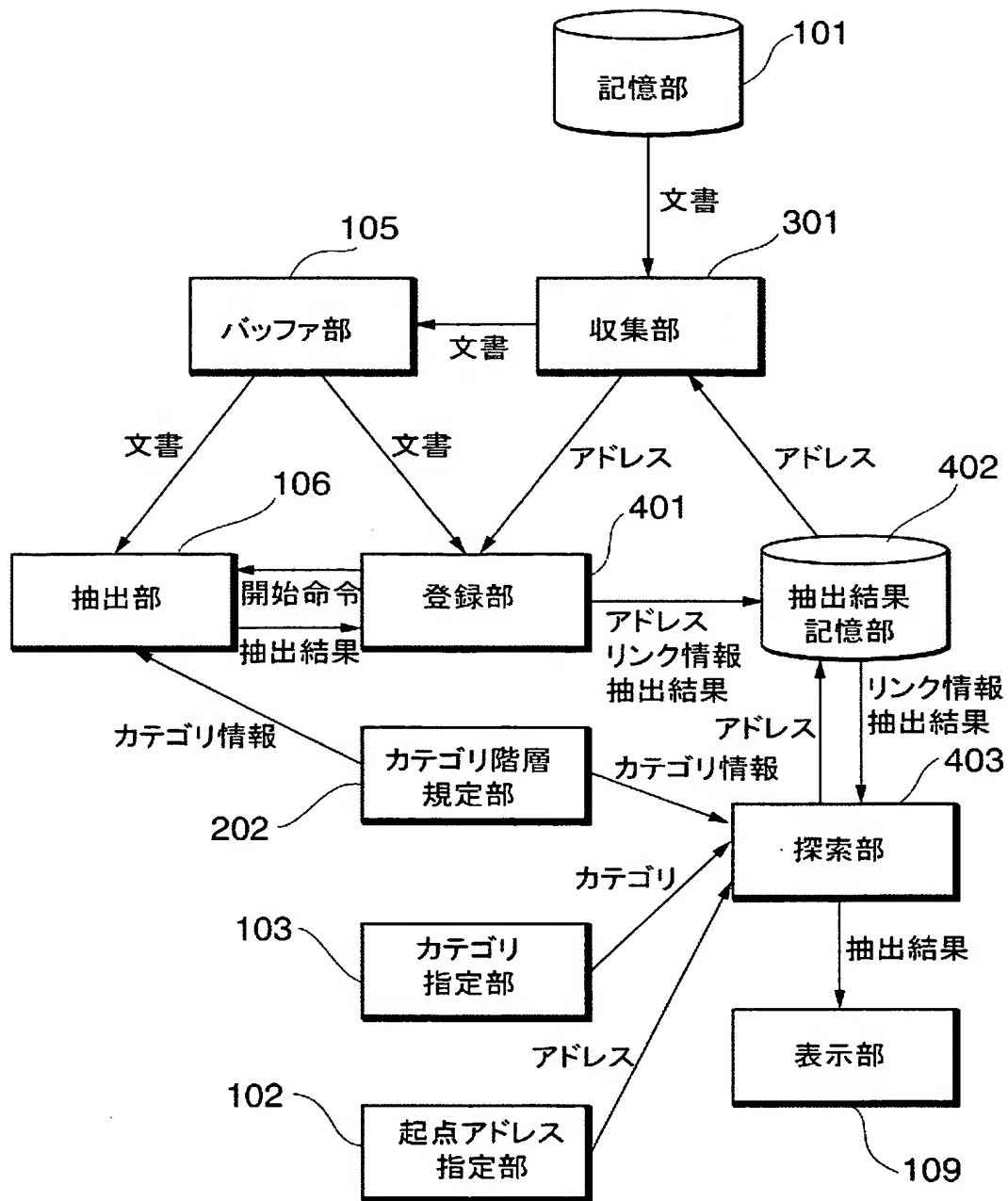
【図17】



具体例3の探索時のフローチャート



【図18】



具体例4の構成図

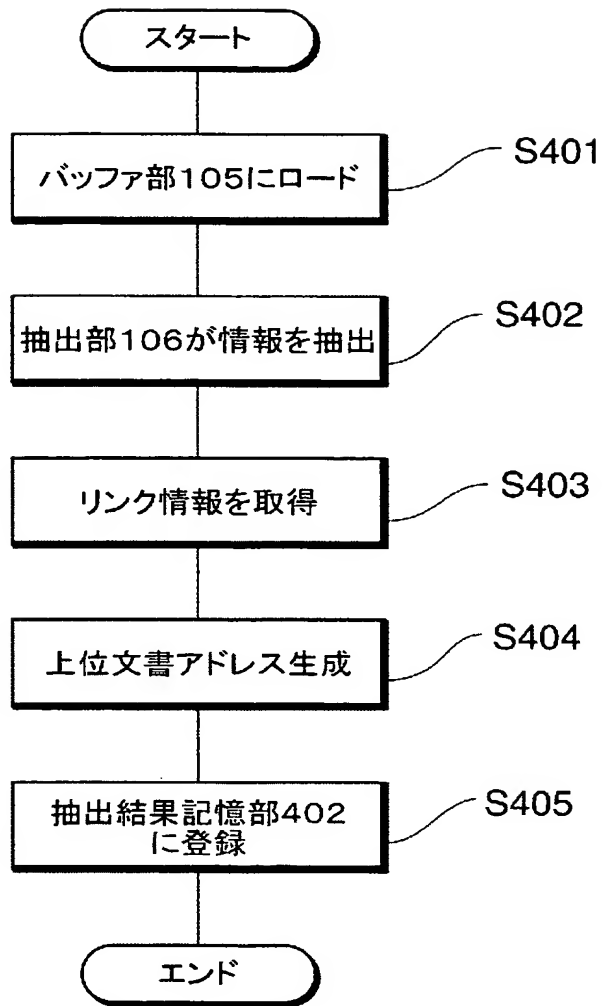
【図19】

	アドレス	抽出結果		リンク情報		上位文書
		カテゴリ	抽出結果	リンク先	リンク元	
1	index.html	組織名(学科名)	情報工学科	shokai.html map.html		※1
2	shokai.html	組織名(研究室名) 組織名(研究室名) 組織名(研究室名)	秋山研究室 井上研究室 遠藤研究室	lab/01.html lab/02.html lab/03.html	index.html	※1
3	map.html	住所	大阪府…		index.html	※1
4	lab/01.html	組織名(研究室名) 人名 :	秋山研究室 秋山敏彦 :		shokai.html	index.html shokai.html
5	lab/02.html	組織名(研究室名) 人名 :	井上研究室 井上宗男 :		shokai.html	index.html shokai.html
6	lab/03.html	組織名(研究室名) 人名 :	遠藤研究室 遠藤豆太郎 :		shokai.html	index.html shokai.html
	:					

※1:shousei.ac.jp/kgb/index.html

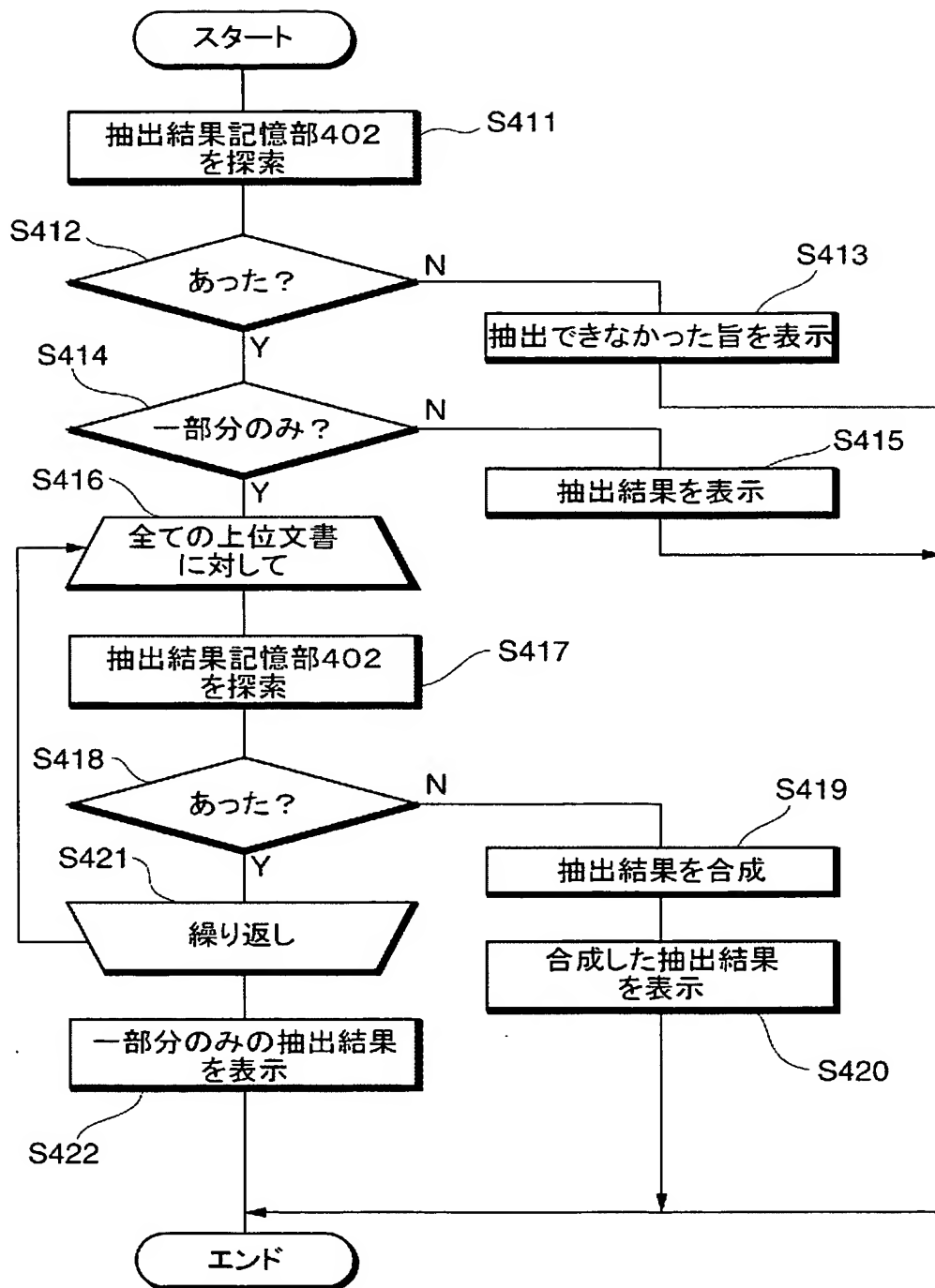
具体例4の抽出結果記憶部の内部データの説明図

【図 20】



具体例4の登録時のフローチャート

【図 21】



具体例4の探索時のフローチャート

【書類名】 要約書

【要約】

【課題】 ハイパーテキスト形式の文書であっても情報抽出を的確に行うことのできる情報抽出装置を得る。

【解決手段】 情報を抽出する場合、起点アドレス指定部 102 で起点となる文書のアドレスを指定する。また、最大リンク深度指定部 104 で最大リンク深度を指定する。抽出部 106 は、起点として指定された対象文書から情報を抽出すると共に、対象文書から情報を抽出できなかった場合は、文書のアドレスに基づいて対象文書のリンク先文書から最大リンク深度の範囲内で情報を抽出する。

【選択図】 図 1

認定・付加情報

特許出願の番号	特願 2 0 0 3 - 0 9 8 1 6 5
受付番号	5 0 3 0 0 5 4 3 2 3 7
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 5 年 4 月 2 日

<認定情報・付加情報>

【提出日】	平成15年 4月 1日
-------	-------------

次頁無

特願 2 0 0 3 - 0 9 8 1 6 5

出 願 人 履 歷 情 報

識別番号

[ 0 0 0 0 0 0 2 9 5 ]

1. 変更年月日

1 9 9 0 年 8 月 2 2 日

[変更理由]

新規登録

住 所

東京都港区虎ノ門1丁目7番12号

氏 名

沖電気工業株式会社